

*f*oisonnement de documents de tailles et de qualités hétéroclites, démocratisation de l'édition, relations hypertextuelles faisant fi des hiérarchisations verticales de l'information, apparition d'annuaires et de classifications « maison », avènement de l'arrogante recherche « plein texte »... parmi les bouleversements suscités par Internet, celui qui touche aux langages documentaires est considérable. Le point en trois volets : évolution, état des lieux, perspectives pour le Web éducatif.

- Parallèlement, l'Éducation nationale s'organise, comme en témoignent les derniers développements de concert de Motbis, qui cherche à faire entrer Internet dans le CDI, et de BCDI Web, qui fait le contraire.
- D'autres outils pour les documentalistes apparaissent : des solutions libres – ou non – de gestion des bases documentaires, un navigateur équipé d'outils pédagogiques...

# L'évolution des langages

## 1. L'héritage classi

Bruno Menon

ENSEIGNANT EN SCIENCES DE L'INFORMATION

*Malgré les moteurs de recherche sémantiques, les classifications originales proposées par les annuaires web et les taxonomies d'entreprises pensées en fonction du client, il existe une continuité entre les langages documentaires classiques et ceux induits par Internet.*

# documentaires

que



« Les thésaurus offrent une réponse à la dématérialisation de la notice, qui rendait hasardeuse la recherche séquentielle (systématique ou alphabétique)... »

**A** lors que les pratiques de l'information-documentation ont connu ces dernières années de profondes transformations, et que l'on se pose régulièrement la question de l'avenir des langages documentaires<sup>1</sup>, Jacques Chaumier pouvait affirmer, en l'an 2000 : « *Les langages documentaires n'ont pas bougé*<sup>2</sup>. » Cette remarque est doublement fondée.

Tout d'abord, aucune architecture radicalement nouvelle de langage documentaire n'est apparue depuis environ un demi-siècle : la plus récente innovation en la matière remonte à 1958, avec l'apparition des thésaurus<sup>3</sup>. D'autre part, le développement des langages documentaires au cours du XX<sup>e</sup> siècle est surtout le fait d'une diversification, sans remise en cause des modèles établis. S'il est peu fréquent que de nouvelles formes voient le jour, il est plus rare encore qu'elles s'éteignent. Les trois souches principales de langages documentaires – classifications, listes de vedettes-matières et thésaurus – sont fermement enracinées dans les pratiques professionnelles, et les instruments qui en sont issus, abondamment utilisés et constamment maintenus.

Pour autant, l'immobilisme n'est qu'apparent. On verra toutefois que les nouveautés ne sont pas à rechercher dans l'univers habituel des centres d'information (bibliothèques, bases de données, centres de documentation), mais dans le monde des entreprises, et dans la nébuleuse Internet.

## Les facteurs de l'évolution

Dans le passé, l'évolution des langages documentaires s'est faite en fonction de mutations culturelles ou technologiques qui conduisaient à repenser la question de l'organisation des connaissances et de l'accès à l'information.

Ainsi les classifications encyclopédiques des bibliothèques résultent-elles du foisonnement éditorial de la fin du XIX<sup>e</sup> siècle<sup>4</sup>, qui provoque un accroissement considérable des collections<sup>5</sup>.

Les listes de vedettes-matières réagissent quant à elles à une accélération de la circulation des savoirs : l'établissement de catalogues imprimés, avec leur indexation par matières, permet la diffusion des informations bibliographiques ; la publication des travaux scientifiques se fait de plus en plus sous forme d'articles de périodiques, d'où un découplage entre unité documentaire et unité physique qui limite le pouvoir du classement et incite à d'autres formes d'indexation<sup>6</sup>.

Avec l'apparition de l'informatique, les thésaurus offrent une réponse à la dématérialisation de la notice, qui rendait hasardeuse la recherche séquentielle (systématique ou alphabétique), et aux nouvelles possibilités techniques ; la logique de recherche booléenne qui les sous-tend est remarquablement adaptée aux opérations de base effectuées par l'ordinateur, et leur succès immédiat ne s'est plus démenti.

Les initiatives plus ou moins concurrentes des thésaurus ont moins bien réussi. Les langages syntagmatiques (par exemple Codoc, Syntol ou Precis) comportent non seulement un lexique,

1. Voir, par exemple, la journée d'étude de l'ADBS, *Du thésaurus au Web sémantique. Les langages documentaires ont-ils encore un avenir ?* 11 avril 2002.

2. Entretien paru dans *Archimag* n° 139, nov. 2000.

3. Le terme *thésaurus* semble être utilisé pour la première fois avec son acception actuelle dans : Needham (R.M.) et Joyce (T.) *The thesaurus approach to information retrieval*. American Documentation, 9 (3), 1958, 192 – 197.

4. Au Royaume-Uni, par exemple, on passe de 600 nouveaux titres annuels vers 1825 à 6000 à la fin du siècle (source : *Encyclopaedia Britannica*).

5. Dans la préface à la première édition de sa classification, Melvil Dewey mentionne à plusieurs reprises le gain de place comme l'un des avantages majeurs de son système.

6. Aujourd'hui quasiment réservées au catalogue des monographies, les vedettes-matières étaient initialement conçues pour traiter toutes formes de production éditoriale, articles de périodiques compris.



**Yahoo! fait de la catégorie *Développement durable* à la fois une subdivision de *Développement économique* et de *Protection et préservation*.**

comme les thésaurus, mais aussi une syntaxe qui leur donne une expressivité et une précision accrues. Cela au prix d'une complexité qui les a disqualifiés au regard de la pure combinatoire illustrée par les thésaurus. Ces langages sont inspirés de l'approche des classifications à facettes, lesquelles ne se sont guère répandues, pour des raisons similaires.

De ces observations, on peut déduire qu'un langage documentaire prospère s'il est adapté à sa fonction et constitue la solution la plus économique à un problème nouveau posé par un bouleversement donné de l'environnement informationnel.

### Une disparition annoncée

L'essor d'Internet constitue justement une mutation technologique et culturelle de première grandeur, et l'on pouvait s'attendre à ce qu'il engendre un renouveau des langages documentaires et favorise l'apparition de pratiques d'indexation en adéquation avec le nouvel ordre mondial de l'information. Mais l'engouement considérable pour les moteurs de recherche sur le texte intégral a conduit à prédire un peu hâtivement la fin des langages documentaires et des pratiques d'indexation classiques, au profit de l'automatisme. Les procédés d'indexation automatique de type sta-

tistique ou statistico-linguistique font l'objet de recherches actives et permettent d'obtenir des résultats parfois impressionnants<sup>7</sup>. Ils comportent cependant des imperfections et des inconvénients : l'apprentissage statistique est finalement un art plus qu'une science, en particulier en ce qui concerne le choix des corpus d'entraînement ; les requêtes qu'ils traitent doivent comporter un matériel lexical relativement abondant pour être interprétées ; les réponses qu'ils proposent sont parfois assez surprenantes, du fait des pondérations attribuées aux mots du texte et/ou de la requête ; enfin, les efforts pour arriver à un traitement correct des phénomènes de synonymie, d'ambiguïté et des compositions lexicales n'ont à ce jour pas abouti.

Mais surtout, pour ces raisons ou pour d'autres, ces systèmes ne sont pas inclus dans les moteurs d'indexation et de recherche du marché, ou le sont sous une forme trop rudimentaire pour être efficace. C'est la plupart du temps un mode de recherche fondé sur les opérateurs booléens qui est présent, même si ces opérateurs sont souvent implicites. Ces outils permettent donc tout au plus de vérifier la présence ou la coprésence dans une ressource d'un ou plusieurs mots : c'est presque par hasard qu'ils peuvent être utilisés pour effectuer une recherche thématique.

Ces techniques ne se substituent pas à l'indexation à base de langages documentaires. Elles ne sont ni des prédatrices ni des concurrentes des langages documentaires, mais devraient occuper une niche écologique créée par la prolifération de l'information électronique : le traitement de documents dont l'intérêt économique ou scientifique ne justifie pas un traitement documentaire classique, forcément coûteux. Elles pourraient aussi entretenir une sorte de relation symbiotique avec des systèmes à base de classifications ou de langages contrôlés. C'est du reste la coexistence ou la coopération d'un moteur de recherche avec un accès par système de classification qui constitue aujourd'hui le dispositif le plus répandu sur le World Wide Web.

### Des classifications encyclopédiques aux répertoires web

Un siècle d'utilisation des classifications encyclopédiques a démontré la puissance d'un mode d'organisation de l'information fondé sur une hiérarchie de classes emboîtées. Les systèmes de catégories employés par les grands répertoires web héritent, consciemment ou pas, de cette tradition. Le principe de navigation – fondateur du Web – et le caractère intuitif d'un parcours du général au particulier dans une hiérarchie en font des instruments là encore très bien adaptés à

7. Voir, par exemple, les comptes rendus des conférences TREC (*Text Retrieval Conference*) sur l'évaluation des dispositifs de recherche d'information, à l'URL <http://trec.nist.gov>

leur fonction. Outre Yahoo! et l'*Open Directory Project*, on ne compte plus les sites « portails » qui ont adopté cette configuration, en créant et faisant évoluer chaque fois leur propre schéma de classification. Les classifications encyclopédiques existantes, dont la couverture paraît pourtant suffisante pour la description des ressources sur Internet<sup>8</sup>, n'ont pas été retenues pour servir de canevas à ces répertoires. La classification Dewey a fait l'objet de tentatives dans ce sens, notamment pour recenser des sites à contenu informatif dense, mais les répertoires qui en font usage sont de taille restreinte et restent assez marginaux<sup>9</sup>.

Les arguments ne manquent pas pour justifier l'adoption de schémas « maison », qui échappent aux critiques habituelles à l'encontre des classifications de bibliothèques : la relative complexité de la construction des indices est évitée, puisqu'on se passe de toute base de notation ; le manque d'hospitalité d'une classification décimale, strictement limitée à dix subdivisions par classe<sup>10</sup> est par là même résolu ; la rigidité de la monohiérarchie, qui conduit à des choix cornéliens, est dédaignée au profit d'une polyhiérarchie rendue transparente par le jeu des liens hypertextuels<sup>11</sup> ; enfin, la lenteur des évolutions et la pesanteur du processus de gestion effectué dans un cadre institutionnel ne sont plus des obstacles, et la classification est mise à jour quotidiennement au besoin, de façon à cerner au plus près les préoccupations de ses usagers.

Malgré les défauts que présentent inévitablement ces instruments sur le plan conceptuel – en particulier l'hétérogénéité des principes de division présidant à l'affinement d'une classe<sup>12</sup> et le nombre sans doute excessif de catégories<sup>13</sup> –, ils sont à l'origine d'un regain d'intérêt pour les classifications et démontrent la viabilité d'un traitement intellectuel de l'information, même à l'heure d'Internet.

### Des classifications spécialisées aux taxonomies d'entreprise

Avec l'implication de nombre d'entreprises dans le « *e-business* », c'est-à-dire toute activité commerciale ou de communication interne et externe exercée au moyen de l'infrastructure d'Internet, ou encore dans des processus de capitalisation des connaissances, la nécessité d'une organisation rationnelle de l'information devient impérieuse, bien au-delà des problématiques bibliographiques. On peut regrouper sous le vocable

**« Les langages documentaires sortent ainsi de la sphère scientifique et technique pour devenir des auxiliaires de l'économie marchande et des outils de management à part entière. »**

« taxonomies d'entreprise » les systèmes d'organisation des connaissances mis au point dans ce contexte. Mais pourquoi pas « thésaurus » ou « classifications » ? Tout d'abord du fait d'une volonté plus ou moins explicite d'éviter les connotations jugées peu flatteuses de ces termes liés à des pratiques trop anciennes. Mais aussi pour échapper aux normes<sup>14</sup> et aux traditions qui gouvernent l'architecture des langages documentaires classiques. De nature avant tout classificatoire, ces instruments servent bien entendu au classement et à la présentation des informations ; mais ils sont conçus pour refléter et projeter une conception des métiers, des savoir-faire, des modes de fonctionnement de l'organisation, bref une culture d'entreprise. À ce titre, ils peuvent aussi intégrer une composante terminologique avec définitions et contrôle du vocabulaire, et servir à la production de métadonnées. Les modèles possibles pour ces taxonomies sont très variés : il est convenu qu'elles doivent au moins présenter une structure hiérarchique, mais on peut y adjoindre différents types de relations sémantiques, adopter une organisation par facettes, etc.<sup>15</sup>. Liberté et diversité structurelles sont donc la règle, mais les promoteurs des taxonomies d'entreprise s'accordent sur l'importance d'une méthodologie de conception fortement ancrée dans le contexte entrepreneurial (clients, produits, processus). Les langages documentaires sortent ainsi de la sphère scientifique et technique pour devenir des auxiliaires de l'économie marchande et des outils de management à part entière.

### Des thésaurus aux ontologies

Le Web sémantique a pour objectif de permettre une exploitation automatique des ressources disponibles en ligne, afin de proposer des fonctions nouvelles ou de meilleure qualité : moteurs de recherche intelligents, combinaison de différents services pour mener à bien des opérations plus ou moins complexes de la vie quotidienne ou professionnelle, navigation sémantique. Il repose sur des formalismes normalisés, sur la description des ressources au moyen de métadonnées, et sur la codification dans des ontologies des connaissances nécessaires à la création et à l'exploitation de ces métadonnées.

Les analogies entre ontologies et thésaurus sont patentées : en marge de leur rôle dans l'indexation et le repérage de l'information, ces derniers servent également à recenser et à structu-

8. Voir, par exemple : Vazine-Goetz, D. *Using library classification schemes for Internet resources (Position Paper)*. *Proceedings of the OCLC Internet Cataloging Colloquium*, San Antonio, Texas, 19 janvier 1996. Dublin, Ohio : OCLC.

9. On trouvera à l'URL [www.public.iastate.edu/%7E/CYBERSTACKS/CTW.htm](http://www.public.iastate.edu/%7E/CYBERSTACKS/CTW.htm) une liste de ces répertoires.

10. Sauf à utiliser le redoutable artefact de l'octave (réserver le chiffre 9 pour introduire une nouvelle série de subdivisions).

11. Yahoo! fait ainsi de la catégorie *Développement durable* à la fois une subdivision de *Développement économique et de Protection et préservation*.

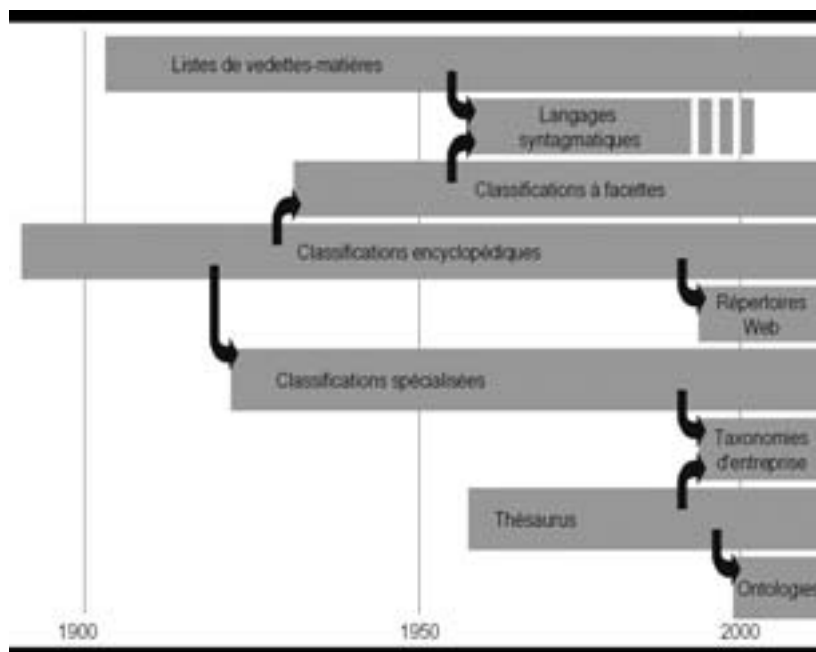
12. On trouve dans Yahoo!, dès le premier niveau, des catégories thématiques (*Sciences et technologies*), géographiques (*Classement régional*), par activité (*Diversification*) et par forme de documents (*Références et annuaires*).

13. Plus de 590 000 pour *Open Directory Project* (source <http://dmoz.org/>).

14. Les normes en matière de thésaurus, peut-être trop anciennes, ne sont pourtant guère contraignantes ; elles fournissent un socle structurel et méthodologique à partir duquel de nombreuses variations sont possibles, et attestées.

15. *eBay* est un exemple de taxonomie marchande universelle dont on peut apprécier la souplesse en consultant ses diverses déclinaisons nationales (France, États-Unis, Allemagne, Royaume-Uni, etc.) : variété des principes de division, présence de facettes à certains points de la hiérarchie.

**« L'engouement considérable pour les moteurs de recherche sur le texte intégral a conduit à prédire un peu hâtivement la fin des langages documentaires et des pratiques d'indexation classiques. »**



**Évolution des langages documentaires depuis 1900.**

#### Références bibliographiques

- Contat, Odile, **Langages documentaires et nouvelles technologies: l'avenir des langages et leur positionnement au cœur des systèmes d'informations dans le contexte de la presse**. Diplôme d'études supérieures spécialisées (DESS). Bayard (centre de documentation), Institut national des techniques de la documentation du CNAM, 5 novembre 2003. [http://memic.ccsd.cnrs.fr/documents/archives0/00/00/00/57/mem\\_00000057\\_00/mem\\_00000057.pdf](http://memic.ccsd.cnrs.fr/documents/archives0/00/00/00/57/mem_00000057_00/mem_00000057.pdf)
- Deniau, Alina, **Moteurs de recherche et restitution de l'information dans les grandes entreprises: l'exemple du portail Cyberthèque de la Direction des systèmes d'information de la Société générale**. Diplôme d'études supérieures spécialisées (DESS), Société générale, Institut national des techniques de la documentation du CNAM, 25 novembre 2003. [http://memic.ccsd.cnrs.fr/documents/archives0/00/00/00/13/mem\\_00000013\\_00/mem\\_00000013.pdf](http://memic.ccsd.cnrs.fr/documents/archives0/00/00/00/13/mem_00000013_00/mem_00000013.pdf)
- Fayet-Scribe, Sylvie (ed.), **Le savoir et ses outils d'accès: Repères historiques**. Solaris. 1997; 4. Rennes: Presses universitaires de Rennes. <http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d04/>
- Maniez, Jacques, **Actualité des langages documentaires: fondements théoriques de la recherche d'information**. Paris: ADBS éditions, 2002.
- Weinberg, Bella Hass, **Complexity in Indexing Systems, Abandonment and Failure: Implications For Organizing The Internet**. ASIS 1996 Annual Conference Proceedings. [www.asis.org/annual-96/ElectronicProceeding/weinberg.html](http://www.asis.org/annual-96/ElectronicProceeding/weinberg.html)

rer la terminologie et les concepts d'un métier ou d'une discipline; or la représentation des connaissances dans les ontologies est fondée sur des termes, des concepts et des relations sémantiques. La structuration des concepts en réseau et la normalisation de leur expression constituent des points importants<sup>16</sup>. Dans la mesure où il existe des thésaurus dans des domaines très variés, comportant des milliers de termes pertinents, il est souvent judicieux de les intégrer dans des ontologies<sup>17</sup>, voire d'en faire leur noyau<sup>18</sup>.

Il faut cependant noter que les thésaurus doivent être remaniés et étoffés pour en permettre une exploitation automatisée. Par exemple, il est souvent nécessaire d'insérer dans les ontologies des connaissances sur des personnes, des lieux, ou des produits, d'y incorporer plusieurs langages documentaires, d'y ajouter des relations plus fines que celles que l'on connaît traditionnellement. De plus, les métadonnées ne se limitent pas à la description thématique des ressources: en fonction des applications visées, elles peuvent comporter des informations plus factuelles, des données signalétiques, etc.<sup>19</sup>.

On peut voir dans les ontologies les descendantes surdouées des thésaurus, encore au berceau, mais susceptibles une fois parvenues à maturité de devenir des outils vraiment universels de caractérisation de l'information.

#### Une généalogie des langages documentaires

Internet a donc bien suscité une nouvelle génération de systèmes d'organisation des connaissances, dans lesquels on peut reconnaître la plupart des composantes des langages documentaires traditionnels, et qui en sont issus en ligne plus ou moins directe. L'héritage n'est pas toujours revendiqué, mais l'examen des parentés structurelles permet d'évoquer cette généalogie.

Que l'on choisisse ou non de les baptiser « langages documentaires », ces systèmes témoignent de la renaissance des pratiques de description des contenus et d'organisation des connaissances. Le pragmatisme et le syncrétisme qui marquent souvent leur conception montrent que les oppositions entre langages classificatoires et combinatoires, précoordonnés et postcoordonnés, énumératifs et à facettes, peuvent être dépassées, dès lors qu'un usage avisé des nouvelles technologies autorise une souplesse inédite et permet de combiner ce que chacune de ces structures a de plus efficace. C'est toujours au final l'adéquation de ces langages à leurs conditions d'usage qui en garantit la pérennité. ●

16. Elles ne doivent pas masquer les spécificités de chacun de ces instruments, qui dérivent de vocations dissemblables: les thésaurus sont adaptés à leur rôle d'outils de médiation documentaire, les ontologies doivent servir à la représentation de multiples aspects des ressources numériques.

17. Comme c'est le cas pour le projet *Hi-Touch*, dans le domaine du tourisme, qui inclut dans son ontologie le thésaurus de l'OMT (Organisation mondiale du tourisme).

18. On peut citer l'exemple du projet AOS (*Agricultural Ontology Service Project*) mené par l'OAA

(Organisation des Nations unies pour l'alimentation et l'agriculture): [www.fao.org/agris/aos/About.htm](http://www.fao.org/agris/aos/About.htm)

19. Une ontologie serait à une application de Web sémantique ce que l'ensemble des fichiers d'autorité (matières, auteurs, titres) est à un catalogue de bibliothèque.